AlloSpatial: Agentic Harness Framework for Spatial Reasoning in Foundation Models

Shouwei Ruan[1] **Bin Wang**[2] **Zhenyu Wu**[1] **Qihui Zhu**[1] **Yuxiang Zhang**[2] **Jingzhi Li**[3]
Yubin Wang[2] **Xingxing Wei**[1*]

[1] Institute of Artificial Intelligence, State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2] Huawei Noah's Ark Lab
[3] University of Science and Technology Beijing

Abstract

Multimodal Foundation Models (MFMs) have made substantial progress, yet remain fragile in spatial reasoning over the physical world. A key bottleneck lies in their inability to transform local egocentric observations into a global allocentric spatial representation. To address this, we propose **AlloSpatial**, an agentic framework for allocentric spatial cognition in foundation models. AlloSpatial introduces **World2Mind**, a plug-and-play cognitive mapping sandbox that converts egocentric observations into structured allocentric priors, including Allocentric-Spatial Trees and route maps that support querying object topology, geometric relations, passability, and trajectories. To utilize these priors reliably under noisy reconstruction and ambiguous visual evidence, AlloSpatial introduces a **Spatial Reasoning Harness** for tool-use judgment, modality-decoupled cue collection, and geometry-semantic arbitration. We further internalize this process in Qwen3-VL through cold-start reinforcement learning with a harness-gated trajectory-level reward. Experiments on VSI-Bench and MindCube show that AlloSpatial improves proprietary models by 5%–18% in a training-free setting, while ASTs alone support strong spatial reasoning even when visual inputs are removed. The trained AlloSpatial agents further outperform larger general-purpose models and competitive spatial baselines, suggesting that structured allocentric representations, active tool use, and verifiable reasoning offer a promising route toward spatially capable foundation models.

1 Introduction

Multimodal Foundation Models (MFMs) [1, 25, 32, 13, 33] have made substantial progress in cross-modal understanding and reasoning. Yet their ability to reason about 3D space in the physical world remains fragile [41, 35, 19, 17, 28]. A central limitation is that current MFMs largely operate over local, transient, and egocentric observations, lacking a mechanism to transform partial perceptual evidence into global, persistent, and queryable mental representations that reliably bridge semantic understanding and geometric relations [35, 33, 42].

Existing approaches have improved spatial reasoning along three paradigms. **Vision-centric methods** post-train MFMs on large-scale 3D-grounded samples, encouraging models to infer depth, size, location, or spatial relations directly from limited visual observations [10, 7, 21, 39]. While effective under in-distribution settings, recent studies [27, 16, 40] indicate that their gains are often coupled to the statistics of the training distribution and may degrade under shifts in scene layout. **Geometry-centric methods** inject explicit spatial signals, such as depth maps, point clouds, or learned 3D

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

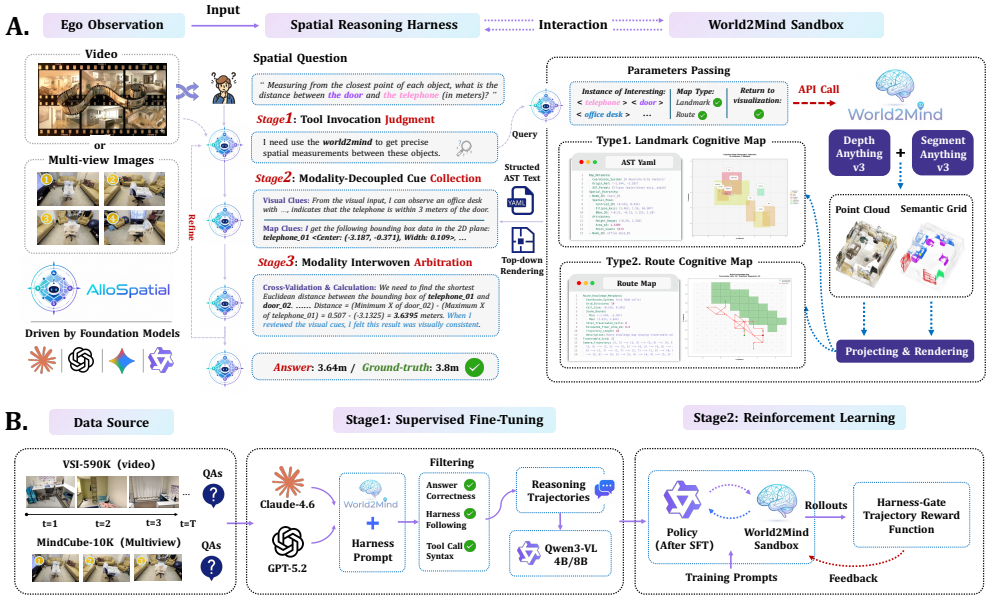


Figure 1: **AlloSpatial inference and training pipeline.** **A**, At inference time, AlloSpatial takes egocentric videos or multi-view observations and follows a three-stage Spatial Reasoning Harness to invoke World2Mind, acquire allocentric spatial knowledge, and arbitrate evidence before answering. **B**, To internalize this harness, we distill and filter high-quality harness-following trajectories from proprietary models for supervised cold start, and further optimize the policy with live World2Mind interaction and a Harness-Gated Trajectory Reward.

representations, to compensate for the weak geometric grounding of 2D visual inputs [11, 8, 23, 38]. Although useful, such signals introduce nontrivial cross-modal alignment challenges [45] and often require geometry-rich paired data and costly training for specialized architectures. More recently, **tool-augmented spatial reasoning** enables models to call tools or task-specific models (e.g., 3D reconstruction, novel-view rendering, depth estimation, or pose estimation) for active evidence acquisition [46, 20, 9]. However, these tools typically return pixel-space observations or low-level geometric measurements, leaving the model to assemble high-level spatial structure through a long, error-prone reasoning chain. As a result, noisy or incomplete tool evidence can be absorbed rather than challenged, leading to incorrect reasoning results.

Indeed, biological intelligence (BI) offers a natural blueprint for overcoming the spatial reasoning bottleneck. Decades of cognitive science research suggest that BI does not passively match each incoming egocentric observation in isolation; instead, it compresses local perceptual experience into stable allocentric cognitive maps [3], supporting mental simulation [24, 31, 2]. This suggests a sharper hypothesis: ***robust spatial reasoning is constrained not merely by coarse 2D perception or limited 3D supervision, but by the absence of an allocentric spatial cognition framework that can be actively invoked, precisely queried, and cross-validated by foundation models.*** Building on this view, we propose **AlloSpatial**, an agentic framework that enables reasoning over structured allocentric spatial representations rather than relying solely on egocentric observations.

As illustrated in Fig. 1(A), at the core of AlloSpatial is **World2Mind**, a plug-and-play cognitive mapping sandbox that transforms egocentric videos or images into queryable allocentric spatial priors. Given semantic categories specified by the agent, World2Mind integrates a robust semantic-geometry alignment pipeline to construct a sparse semantic point cloud. It then distills this point cloud into two complementary cognitive maps: a **Landmark Cognitive Map** for object-centric topological reasoning and a **Route Cognitive Map** for traversability, passability, and trajectory reasoning. Its central representation is the proposed **Allocentric-Spatial Tree (AST)**, a directed acyclic graph whose nodes correspond to stable environmental landmarks and whose attributes encode centroid, footprint, principal axes, orientation, height range, and hierarchical containment. Unlike grid maps that discard object identity [41, 35, 30] or abstract semantic graphs that omit metric geometry [15, 29], AST compresses noisy 3D reconstruction into compact and structured spatial memory.

However, structured spatial priors alone do not guarantee reliable reasoning, especially when reconstruction drift and perception errors often occur. AlloSpatial therefore introduces a three-stage **Spatial Reasoning Harness** to regulate how models invoke tools, collect evidence, and arbitrate across modalities. The harness first determines whether a question truly requires cognitive mapping, then decouples evidence collection from raw visual inputs, AST-structured text, and optional top-down visualization maps produced by World2Mind, and finally performs geometry-semantics interleaved reasoning to identify conflicts and cross-validate the final answer. To internalize the reasoning harness, we train an AlloSpatial agent instantiated from Qwen3-VL (see Fig. 1(B)). We first use World2Mind within the harness prompt to distill high-quality trajectories from frontier proprietary models, and then apply supervised cold-start fine-tuning to bootstrap tool invocation and reasoning structure. The agent is subsequently optimized with Group Sequence Policy Optimization (GSPO) [48]. Since RL over long tool-use trajectories is vulnerable to reward hacking and redundant tool calls, we introduce a **Harness-Gated Trajectory Reward (HGTR)** that evaluates the trajectory as a whole rather than isolated actions. HGTR jointly accounts for answer correctness, harness compliance, tool-use validity, and response efficiency, while enforcing two key gates: answer accuracy is credited only when the trajectory follows the required reasoning structure, and tool-use rewards are granted only when valid tool calls contribute to a correct answer. This training strategy stabilizes optimization and enables the agent to acquire efficient and deliberate allocentric spatial reasoning behaviors.

We evaluate AlloSpatial on VSI-Bench [41] and MindCube [35], covering various spatial reasoning tasks across egocentric videos and sparse multi-view images. As a training-free plug-in, World2Mind combined with the Spatial Reasoning Harness consistently improves frontier commercial models, including GPT-5.2, Claude-4.6, and Gemini-3, with overall gains of approximately 5%-18%. Under a "blind" setting in which visual inputs are removed, ASTs alone achieve strong 3D reasoning performance, suggesting that high-quality allocentric priors can elicit spatial mental simulation even without direct visual evidence. After cold-start RL, AlloSpatial agents instantiated from Qwen3-VL surpass larger frontier models and competitive methods across multiple tasks. In summary, our study indicates that coupling structured allocentric representations with an agentic reasoning harness offers a promising route toward overcoming the spatial cognition bottleneck of foundation models.

2 Related Work: Spatial Reasoning in Foundation Models

Recent benchmarks have exposed spatial reasoning as a persistent weakness of MFMs. VSI-Bench [41] and VSI-Super [42] evaluate egocentric video-based spatial intelligence. MindCube [35] probes sparse multi-view cognitive mapping, while other benchmarks [19, 28] extend evaluation to broader text-based and multimodal scenarios. Together, these studies suggest that the bottleneck is not merely visual recognition, but the absence of reasoning-ready spatial mental representations.

Vision-centric learning. Vision-centric methods improve spatial reasoning by post-training MFMs with large-scale 3D-grounded supervision. SpatialVLM [7], Cambrian-S [42], and related works [26, 35, 12] construct spatial VQA data from real-world images, videos, or 3D annotations to supervise spatial relation understanding. SpatialReasoner [21] and Spatial-MLLM [39] further introduce intermediate 3D representations or explicit reasoning traces to structure spatial inference. These methods show that spatially grounded supervision can improve benchmark performance. Still, their gains may remain coupled to training-distribution statistics and degrade under shifts in scene layout, viewpoint, or object composition. Recent analyses further suggest that MFMs can exploit semantic shortcuts rather than acquire robust spatial cognition [27, 16, 40].

Geometry-centric learning. Geometry-centric methods compensate for the weak geometric grounding of 2D inputs by injecting explicit spatial signals. SpatialBot [4] uses RGB-D inputs and depth-centered QA data to improve metric spatial understanding. MM-Spatial [11] and SD-VLM [8] study how depth maps, multi-view observations, and depth-encoded visual features affect 3D reasoning, while N3D-VLM [38] integrates native 3D grounding into a unified MFM. Other works explore point-cloud-enhanced LLMs or reasoning-based segmentation for localization and scene understanding [23, 45]. Although explicit geometry provides useful spatial cues, directly conditioning MFMs on 3D modalities introduces cross-modal alignment challenges and often requires geometry-rich paired data, specialized architectures, and costly training.

Tool-augmented learning. Tool-augmented methods equip MFMs with external modules for active evidence acquisition. Think3D [46] and pySpatial [20] invoke 3D reconstruction, camera pose

estimation, or novel-view rendering for interactive spatial exploration. SpaceTools [9] studies how models coordinate depth, segmentation, and pose tools through reinforcement learning, while SpatialDreamer [5] uses tool hints or visual imagination as intermediate evidence. These approaches mark an important shift from passive perception to active tool use. However, most remain observation-centric: tools typically return rendered views, masks, depth maps, or low-level measurements, leaving the model to assemble high-level spatial structure through long and error-prone reasoning chains. In contrast, AlloSpatial converts noisy 3D reconstructions into **compact allocentric representations** and couples them with a Spatial Reasoning Harness for tool invocation, modality-decoupled evidence collection, and geometry–semantics arbitration. This enables models to reason over structured allocentric spatial memory rather than isolated egocentric observations or low-level geometric cues.

3 Methodology

3.1 Problem Formulation

Given an egocentric observation sequence $\mathcal{I} = \{I_t\}_{t=1}^{T}$ and a spatial question q, our goal is to produce a final answer $\hat{a}$ together with a verifiable spatial reasoning process. We formulate AlloSpatial as a harness-guided tool-using agent $(\pi_\theta, \mathcal{W}, \mathcal{H})$, where π_θ denotes the foundation-model policy, $\mathcal{W}$ is the proposed World2Mind cognitive mapping sandbox, and $\mathcal{H}$ is the Spatial Reasoning Harness. Given $(q, \mathcal{I})$, the agent generates a multi-turn trajectory:

$$\tau = \big(\mathbf{h}_0,\, u_0,\, o_0,\, \mathbf{h}_1,\, u_1,\, o_1,\, \ldots,\, \mathbf{h}_K,\, \hat{a}\big), \tag{1}$$

where $\mathbf{h}_k$ is the dialogue history at round k, and $u_k \sim \pi_\theta(\cdot \mid \mathbf{h}_k)$ is either a natural-language reasoning step or a structured tool call. If u_k is a valid tool call, World2Mind returns an allocentric spatial observation $o_k = \mathcal{W}(u_k; \mathcal{I})$; otherwise, $o_k = \varnothing$. The final answer $\hat{a}$ is emitted within a predefined answer tag for reliable parsing and evaluation.

The harness $\mathcal{H}$ constrains the trajectory by specifying the required evidence channels, their ordering, and the cross-modal arbitration steps before final prediction. It therefore turns external allocentric priors into executable and inspectable reasoning evidence, rather than treating tool outputs as unverified context. For a training-free plug-in setting, $\mathcal{H}$ is instantiated via prompting. While in the training process, $\mathcal{H}$ defines the trajectory structure optimized during RL, as described in Sec. 3.4. Thus, $\mathcal{H}$ serves both as an inference-time protocol and as a training-time inductive bias.

3.2 World2Mind: Allocentric Cognitive Mapping Sandbox

Rather than reconstructing an exhaustive scene model, World2Mind exposes a query-conditioned cognitive mapping interface. Given an egocentric observation sequence $\mathcal{I}$, an open-vocabulary category set $\mathcal{C}$, a requested knowledge type $k \in \{\textit{landmark}, \textit{route}, \textit{both}\}$, a footprint format $f \in \{\textit{rectangle}, \textit{ellipse}\}$, and a scene type $s \in \{\textit{indoor}, \textit{outdoor}\}$, World2Mind returns $\mathcal{M} = \mathcal{W}(\mathcal{I}; \mathcal{C}, k, f, s)$, where $\mathcal{M}$ denotes the structured allocentric spatial knowledge.

3.2.1 Geometry-Semantic Alignment Pipeline

We first align geometry and semantics across the egocentric sequence. For each frame $I_t \in \mathbb{R}^{H\times W}$, we estimates a dense depth map $D_t \in \mathbb{R}^{H\times W}$ and camera pose $T_t \in SE(3)$ using monocular geometry models [18, 34], and extracts open-vocabulary semantic masks $\{M_t^c\}_{c\in\mathcal{C}}$ with SAM 3 [6]. To suppress unreliable geometry near object boundaries, textureless regions, and poorly reconstructed views, we apply a two-level confidence filter to the predicted confidence map $\Gamma_t \in [0,1]^{H\times W}$:

$$V_t(u,v) = \mathbf{1}[\Gamma_t(u,v) > \tau_{\text{px}}] \cdot \mathbf{1}\big[\bar{\Gamma}_t > \tau_{\text{frm}}\big], \quad \bar{\Gamma}_t = \frac{1}{HW}\sum_{u,v}\Gamma_t(u,v), \tag{2}$$

where (u,v) indexes an image pixel, $V_t(u,v) \in \{0,1\}$ is the validity mask, and τ_{px} and τ_{frm} are the pixel- and frame-level confidence thresholds.

Pixels satisfying both constraints are back-projected into the world coordinate system, where $p_t(u,v)$ denotes the 3D point obtained from $D_t(u,v)$ and T_t. Aggregating valid points, semantic labels, and colors across frames yields a global semantic point cloud $\mathcal{P} = \{(p_i, s_i, \text{rgb}_i)\}_{i=1}^{N}$, where $p_i \in \mathbb{R}^3$, $s_i \in \mathcal{C}$, and N is the number of retained points. To remove sparse outliers caused by

boundary leakage and multi-view misalignment, we compute a K-nearest-neighbor density score $\rho_i = \frac{1}{K}\sum_{j\in\mathcal{N}_K(i)} \|p_i - p_j\|_2^{-1}$ for each point, where $\mathcal{N}_K(i)$ denotes the K nearest neighbors of p_i, and discard points below a category-specific density percentile. The resulting point cloud is sparse but geometrically reliable, serving as the substrate for allocentric map construction.

3.2.2 Allocentric Mapping

World2Mind compresses the filtered $\mathcal{P}$ into complementary allocentric representations.

Landmark mapping with Allocentric-Spatial Tree. For each queried category $c \in \mathcal{C}$, World2Mind applies adaptive DBSCAN to separate object instances. Each instance cluster is represented as a node in the **Allocentric-Spatial Tree (AST)** $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes landmark nodes and $\mathcal{E}$ encodes hierarchical containment or support relations. Unlike abstract scene graphs that omit metric geometry [15, 29] or grid maps that discard object identity [41, 35], AST preserves both object semantics and explicit spatial structure. For each node $v \in \mathcal{V}$, World2Mind projects its supporting points onto the ground plane and fits a compact footprint:

$$\phi(v) = \big(c_x, c_z;\ a, b, \theta;\ h_{\min}, h_{\max};\ A, n\big), \tag{3}$$

where (c_x, c_z) is the top-down centroid, (a, b) are the semi-axes of the fitted ellipse, θ is its orientation, $(h_{\min}, h_{\max})$ denotes the vertical extent, A is the footprint area, and n is the number of supporting points, used as a confidence proxy. Elliptical footprints are used by default for stable coarse occupancy, while an axis-aligned rectangular format is returned when nearest-boundary computation is required. The resulting AST is serialized into `YAML` and included in the structured spatial map $\mathcal{M}$. This representation turns noisy 3D reconstruction into compact, model-readable spatial memory: foundation models can directly parse object centers, sizes, orientations, and containment relations without specialized 3D adapters, while the coarse footprint parameterization remains robust to local reconstruction noise and reflects the compressed nature of cognitive maps [3, 2].

Route mapping. For traversability and path-related queries, World2Mind voxelizes points associated with traversable categories, such as floors, and projects them onto a top-down grid. Each grid cell is labeled as traversable, occupied, or unknown. The camera trajectory $\{T_t\}_{t=1}^{T}$, where T_t denotes the camera pose of frame I_t, is projected into the same coordinate system to encode observed motion history and support route-level reasoning.

Optional visual rendering. Beyond structured text, World2Mind can return allocentric visualizations as auxiliary global observations, including top-down AST layouts, route maps, and semantic segmentation maps. These renderings provide a compact visual summary of the global scene layout and serve as additional evidence during geometry–semantics arbitration.

3.3 Spatial Reasoning Harness

Structured allocentric priors are useful but not self-verifying. Directly conditioning on World2Mind output $\mathcal{M}$ can induce two failure modes: *1) modality lock-in*, where the model prematurely commits to either visual appearance or AST text and ignores conflicting evidence [37, 22]; and *2) over-trust in tool results*, where incomplete or noisy reconstructions are treated as ground truth. To mitigate these failures, we design the **Spatial Reasoning Harness** $\mathcal{H}$ as a cyclic protocol:

$$\mathcal{H}: \quad \textsc{Judge} \rightarrow \textsc{Collect} \rightarrow \textsc{Arbitrate} \rightarrow \{\textsc{Refine}, \textsc{Answer}\}. \tag{4}$$

At each cycle, the agent decides whether additional allocentric evidence is needed, collects evidence through decoupled channels, and arbitrates geometry-semantics conflicts before either refining the query or committing to the final answer.

Stage I: tool invocation judgment. At reasoning cycle r, the agent determines whether the current history $\mathbf{h}_r$ is sufficient or whether World2Mind should be queried. Following the reason-then-act principle of ReAct-style agents [43], it first states a spatial hypothesis and the rationale for tool use:

$$(d_r, \eta_r) = \textsc{Judge}(q, \mathbf{h}_r), \quad d_r \in \{\textsc{Call}, \textsc{Skip}\}, \tag{5}$$

where d_r is the tool-use decision and η_r is the current spatial hypothesis. The agent calls World2Mind only when the question requires metric measurement, route topology, viewpoint transformation, or other allocentric spatial evidence, preserving an initial hypothesis that can later be verified or rejected.

Table 1: Results of AlloSpatial under the training-free setting on the VSI-Bench [41] and MindCube [35] benchmarks (tiny split). For VSI-Bench, we uniformly use 32 input frames.

Models	VSI-Bench									MindCube			
	Overall	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	Overall	Around	Among	Rotation
Proprietary Foundation Models w/o. AlloSpatial													
GPT-5.2	46.7	52.5	34.9	67.5	50.6	42.0	40.7	34.7	51.0	49.9	62.4	45.2	48.5
Claude-4.6-Opus	38.4	46.9	18.5	62.1	26.8	40.0	47.2	34.7	30.6	48.5	58.8	50.7	29.0
Gemini-3-Pro	55.2	47.8	32.1	**71.3**	55.0	54.0	44.8	57.1	79.6	75.1	77.2	68.2	93.0
Proprietary Foundation Models w./ AlloSpatial													
GPT-5.2	54.0 (↑7.3)	47.4 (↓5.1)	33.4 (↓1.5)	63.3 (↓4.2)	52.4 (↑1.8)	**64.0** (↑22.0)	41.1 (↑0.4)	51.0 (↑16.3)	79.6 (↑28.6)	54.6 (↑4.7)	60.4 (↓2.0)	47.7 (↑2.5)	68.0 (↑19.5)
Claude-4.6-Opus	56.0 (↑17.7)	**59.0** (↑12.0)	34.3 (↑15.8)	67.3 (↑5.2)	54.8 (↑28.0)	**64.0** (↑24.0)	62.7 (↑15.6)	**65.3** (↑30.6)	40.8 (↑10.2)	62.9 (↑14.4)	82.4 (↑23.6)	60.8 (↑10.1)	45.0 (↑16.0)
Gemini-3-Pro	**61.0** (↑5.8)	51.8 (↑4.1)	**36.8** (↑4.7)	57.7 (↓13.5)	**62.6** (↑7.6)	62.0 (↑8.0)	**67.7** (↑22.9)	**65.3** (↑8.2)	**83.7** (↑4.1)	**81.6** (↑6.5)	**86.0** (↑8.8)	**75.8** (↑7.6)	**93.5** (↑0.5)

Stage II: modality-decoupled cue collection. When $d_r = \textsc{Call}$, the agent generates query parameters $\xi_r = (\mathcal{C}_r, k_r, f_r, s_r)$ and obtains a structured spatial map $\mathcal{M}_r = \mathcal{W}(\mathcal{I}; \xi_r)$, where $\mathcal{C}_r$ is the queried category set, k_r the requested knowledge type, f_r the footprint format, and s_r the scene type. Evidence is then collected separately from raw visual observations, AST-structured text, and optional top-down maps $\{e_r^{\text{vis}}, e_r^{\text{ast}}, e_r^{\text{map}}\}$. Here, $\mathcal{E}_r$ denotes the evidence set at cycle r, and e_r^{vis}, e_r^{ast}, and e_r^{map} denote visual, textual-allocentric, and rendered-map evidence, respectively. This decoupled collection prevents premature fusion and reduces reliance on a single evidence channel.

Stage III: geometry-semantics interleaved arbitration. The agent then compares visual semantics with geometric evidence from AST and optional maps, identifies conflicts caused by reconstruction drift, missing instances, or false detections, etc., and then decides whether the evidence is sufficient:

$$(\Delta_r, \omega_r, y_r) = \textsc{Arbitrate}(\mathcal{E}_r), \quad y_r \in \{\textsc{Refine}, \textsc{Answer}\}, \tag{6}$$

where Δ_r denotes detected cross-modal conflicts, ω_r represents confidence assignments over evidence channels, and y_r is the next action. If $y_r = \textsc{Refine}$, the agent updates ξ_r and starts another cycle; if $y_r = \textsc{Answer}$, it emits the final response within the `<Answer>` tag. In this way, $\mathcal{H}$ treats external cognitive maps as falsifiable evidence for reasoning rather than as direct pseudo-labels, enabling iterative, verifiable, and trainable allocentric spatial reasoning.

3.4 Internalizing the Spatial Reasoning Harness via Reinforcement Learning

While frontier proprietary models can often follow the Spatial Reasoning Harness $\mathcal{H}$ through prompting, open-weight models [1, 36] do not naturally exhibit stable multi-turn tool-use behavior. In preliminary experiments, they frequently produce malformed tool calls, skip cross-modal arbitration, or degenerate into repetitive interaction. We therefore internalize $\mathcal{H}$ into the open-weight policy π_θ with supervised cold start followed by RL with live World2Mind execution.

Supervised cold start. We first use World2Mind and the harness protocol to distill trajectories from proprietary models. We retain trajectories that are answer-correct, structurally valid, and contain non-trivial cross-modal arbitration. Supervised fine-tuning on these traces bootstraps tool-call syntax, AST and route-map parsing, harness stage ordering, and answer-tag formatting, thereby preparing the policy for subsequent optimization with Group Sequence Policy Optimization (GSPO) [48].

GSPO with Harness-Gated Trajectory Reward. To stabilize reinforcement learning over long tool-use trajectories, we introduce a **Harness-Gated Trajectory Reward (HGTR)** that scores each complete trajectory τ rather than individual tokens. HGTR combines answer correctness, structural compliance, tool-use effectiveness, and response efficiency:

$$R(\tau) = w_{\text{acc}} \widetilde{R}_{\text{acc}}(\tau) + w_{\text{str}} R_{\text{str}}(\tau) + w_{\text{tool}} R_{\text{tool}}(\tau) + w_{\text{len}} R_{\text{len}}(\tau), \tag{7}$$

where $w_{\text{acc}}, w_{\text{str}}, w_{\text{tool}}, w_{\text{len}}$ are reward weights. $R_{\text{str}}(\tau)$ measures harness compliance, including valid tool-call tags, complete stage ordering, explicit arbitration, and the final answer tag. $R_{\text{acc}}(\tau)$ measures answer quality, using exact match for multiple-choice questions and mean relative accuracy

Table 2: Evaluation results on the VSI-Bench benchmark (tiny split). Following the setting of Think3D [46], all models are evaluated with **7** input frames to assess spatial reasoning under limited egocentric observations. †: Results are directly taken from the original paper. Green shades denote the top-1, top-2, and top-3 performance within each metric.

Rank	Models	Overall	Numerical Answer (MRA)					Multiple-Choice Answer (Pass@1 Accuracy)				
			Avg.	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Avg.	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
	Proprietary Foundation Models											
11	GPT-5.2	37.7	36.4	28.6	24.3	52.3	40.6	39.0	42.0	28.4	42.9	42.9
8	Gemini-2.5-Pro	46.2	39.4	34.7	14.3	56.3	52.4	53.0	52.0	45.7	38.8	75.5
9	Gemini-3-Pro	45.2	32.2	29.4	12.6	38.3	48.4	58.3	**56.0**	52.7	**49.0**	75.5
	Open-source Foundation Models											
12	Gemma-4-E4B	33.5	30.1	19.6	24.5	33.1	43.2	36.9	38.0	40.0	30.6	38.8
6	InternVL3.5-4B	48.8	50.9	**77.3**	19.4	60.4	46.4	46.8	38.0	47.1	40.8	61.2
10	Qwen3-VL-4B	45.1	47.7	44.3	34.5	64.4	47.8	42.4	36.0	37.8	24.5	71.4
7	Qwen3-VL-8B	48.5	51.5	50.4	40.2	67.5	48.0	45.5	36.0	44.1	30.6	71.4
3	Qwen3-VL-32B	53.1	55.3	60.6	35.1	71.5	54.0	50.9	46.0	51.4	36.7	69.4
	Specialized Spatial Models											
5	Spatial-MLLM-4B [39]	49.1	53.0	76.5	27.0	57.9	50.6	45.2	44.0	40.9	38.8	57.1
4	Cambrian-S-3B [42]	49.5	51.7	67.1	22.6	**73.8**	43.2	47.3	54.0	37.1	26.5	71.4
–	Think3D-4B† [46]	–	–	–	–	–	–	45.4	44.7	39.0	36.7	61.2
2	AlloSpatial-4B (Ours)	53.5	**56.3**	54.1	**42.3**	68.8	**60.0**	50.8	50.0	59.3	34.7	59.2
1	AlloSpatial-8B (Ours)	**54.2**	47.8	46.1	26.0	65.6	53.4	**60.6**	52.0	**64.1**	46.9	**79.6**

for numerical questions following [41]. To prevent malformed trajectories from receiving reward through accidental correct guesses, HGTR applies a **structure-gated accuracy**:

$$\widetilde{R}_{\mathrm{acc}}(\tau) = R_{\mathrm{acc}}(\tau) \cdot \mathbf{1}[R_{\mathrm{str}}(\tau) \geq \tau_{\mathrm{s}}], \tag{8}$$

where τ_{s} is the threshold and $\mathbf{1}[\cdot]$ is the indicator function. Thus, correctness is credited only when the trajectory satisfies the harness format. HGTR further uses a **correctness-tied tool-use reward**:

$$R_{\mathrm{tool}}(\tau) = \alpha\, \widetilde{R}_{\mathrm{acc}}(\tau)\mathbf{1}[\mathrm{ValidCall}(\tau)] - \gamma\, \max(0,\, n_{\mathcal{W}}(\tau) - n^{\star}), \tag{9}$$

where $\mathrm{ValidCall}(\tau)$ indicates whether World2Mind is invoked with valid syntax and task-relevant arguments, $n_{\mathcal{W}}(\tau)$ is the number of calls, $n^{\star}$ is a soft call budget, and α, γ control the tool reward and overuse penalty. This reward grants tool-use credit only when a valid invocation contributes to a structurally valid and correct answer, while discouraging redundant reconstruction once sufficient allocentric evidence is available. Finally, $R_{\mathrm{len}}(\tau)$ penalizes excessive model-generated tokens while masking out tool-returned ASTs, route maps, and visualization metadata. Overall, HGTR guides the agent toward trajectories that are answer-correct, harness-compliant, tool-efficient, and verifiable.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate AlloSpatial on the official Tiny split of VSI-Bench [41] and MindCube [35], containing 392 and 1,050 questions, respectively. For training, we sample data from VSI-590K and the MindCube training set. The cold-start stage distills multi-turn tool-use trajectories from GPT-5.2 and Claude-4.6-Opus with World2Mind and the Spatial Reasoning Harness, followed by filtering for answer correctness, harness compliance, and valid tool invocation.

Metrics. Following [41, 35], we report pass@1 accuracy for multiple-choice questions and mean relative accuracy (MRA) for numerical questions. The overall score is the unweighted average over all task types within each benchmark. For the training-free setting on VSI-Bench, we uniformly sample up to 32 frames from each video. For the post-trained AlloSpatial agents and other baselines, we follow the limited-observation setting of Think3D [46] and uniformly sample 7 frames, testing performance under sparse visual inputs. We further analyze the effect of frame number in Sec. 4.4.

Baselines. We compare *1) Proprietary models* including GPT-5.2, Claude-4.6-Opus, and Gemini-3-Pro; *2) Open-source models* including Qwen3-VL [1], InternVL3.5 [36], and Gemma-4 [14]; *3) Specialized spatial models* including Spatial-MLLM [39], Cambrian-S [42], and Think3D [46].

Table 3: Evaluation on the MindCube (tiny split).

Rank	Models	Overall	Around	Among	Rotation
2	Gemini-2.5-Pro	57.9	67.2	43.8	**88.5**
5	Gemma-4-E4B	37.3	37.6	38.0	35.0
6	InternVL3.5-4B	36.6	41.6	35.3	34.0
8	Qwen3-VL-4B	28.3	40.4	20.8	35.5
4	Spatial-MLLM-4B [39]	39.5	52.0	36.3	33.5
7	Cambrian-S-3B [42]	33.2	34.4	34.5	28.0
3	Think3D-4B† [46]	44.0	42.5	37.5	42.5
1	AlloSpatial-4B (Ours)	**69.1**	**82.0**	**65.0**	65.5

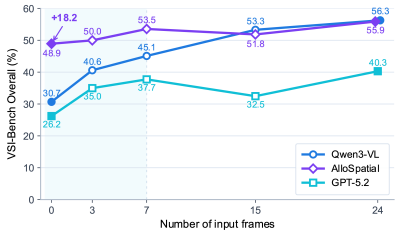


Figure 2: VSI-Bench performance under different input-frame number.

Training details. We instantiate AlloSpatial from Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct [1]. The 4B and 8B agents are trained for 600 and 400 RL steps, respectively, using 4.8K and 2.4K unique prompts, with 8 sampled rollouts per prompt. Detailed hyperparameters, World2Mind service configuration, and computational costs are provided in the Appendices A, B, and E.

4.2 AlloSpatial Improves Proprietary Foundation Models for Spatial Reasoning

We first evaluate AlloSpatial as a *training-free* plug-in for proprietary models by exposing World2Mind $\mathcal{W}$ and the Spatial Reasoning Harness $\mathcal{H}$ through prompting.

Consistent improvement across benchmarks. As shown in Tab. 1, AlloSpatial improves GPT-5.2, Claude-4.6-Opus, and Gemini-3-Pro on VSI-Bench by $+7.3$, $+17.7$, and $+5.8$ overall points, respectively, and on MindCube by $+4.7$, $+14.4$, and $+6.5$ points. The gains are concentrated on tasks that require allocentric structure, such as relative direction, route planning, and viewpoint-dependent rotation, while tasks solvable from local visual evidence show smaller or occasionally negative changes. This pattern suggests that World2Mind is most beneficial when direct egocentric perception is insufficient, and the model must reason over stable spatial relations beyond the observed view. A complete reasoning trace is shown in Fig. 4, and additional analysis for AlloSpatial's reasoning cases are illustrated in Appendix D.

ASTs provide effective allocentric reasoning evidence. We further test whether the gain comes from structured spatial knowledge itself by evaluating VSI-Bench in a text-only "blind" setting [41]. Fig. 3 shows that AST text substantially improves blind spatial reasoning over text-only baselines, especially on object size and route planning. This supports the role of AST as a compact allocentric prior that enables models to reason over spatial structure without directly observing the scene. Together with the full-input results, the blind setting suggests that the critical signal is not merely additional visual context, but the structured spatial organization supplied by World2Mind.

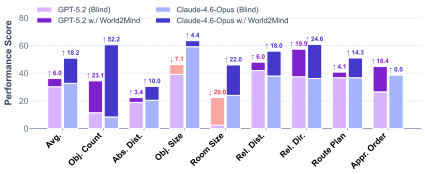


Figure 3: **Text-only spatial reasoning with allocentric priors.** VSI-Bench performance under the "blind" setting, where visual inputs are removed.

4.3 Trained AlloSpatial Agents Outperform General and Spatially Specialized Models

We next compare the trained AlloSpatial agents with proprietary models, open-source MFMs, and specialized spatial reasoning models. As shown in Tab. 2, AlloSpatial-8B achieves the best overall score on VSI-Bench, while AlloSpatial-4B ranks second. Both variants outperform proprietary models such as GPT-5.2, Gemini-2.5-Pro, and Gemini-3-Pro. Notably, AlloSpatial-4B and AlloSpatial-8B also surpass Qwen3-VL-32B, despite using substantially smaller backbones. AlloSpatial also compares favorably with spatially specialized models. On VSI-Bench, AlloSpatial-8B improves over Spatial-MLLM-4B and Cambrian-S-3B by $+5.1$ and $+4.7$ overall points, respectively, while AlloSpatial-4B also exceeds both baselines. This is notable because Spatial-MLLM and Cambrian-S rely on large-scale spatial grounding data, whereas AlloSpatial uses a much smaller set of unique training prompts and acquires spatial competence through structured tool interaction and trajectory-level optimization. These results suggest that allocentric spatial memory and verifiable reasoning can improve data efficiency compared with purely supervision-driven spatial learning.

Considering task-level performance, AlloSpatial shows greater improvements on relational and viewpoint-dependent tasks, such as relative direction, route planning, appearance order, and Mind-

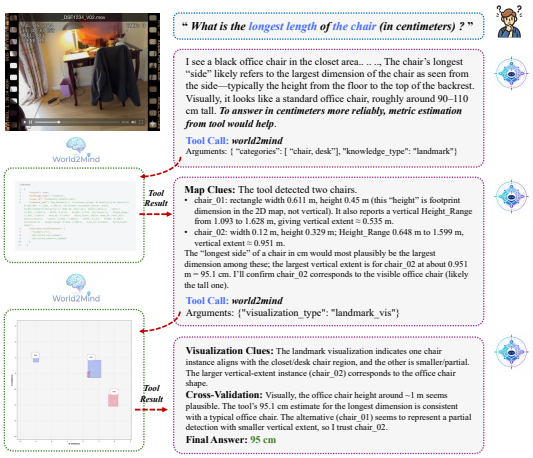


Figure 4: Reasoning trace produced by AlloSpatial.

Table 4: **Ablation results of AlloSpatial-4B.** We compare Qwen3-VL-4B variants using the same training data as QA pairs for SFT and RL, and report AlloSpatial-4B at Stage-1 (SFT cold start) and Stage-2 (RL).

Variants	VSI-Bench	MindCube	Tokens
Qwen3-VL-4B			
Instruct	45.1	28.3	–
Thinking	45.5 (↑0.4)	36.1 (↑7.8)	1064
w./ SFT on QAs	43.1 (↓2.0)	53.9 (↑25.6)	–
w./ RL on QAs	46.2 (↑1.1)	53.0 (↑24.7)	–
AlloSpatial-4B			
Stage-1 (SFT)	38.2 (↓6.9)	52.0 (↑23.7)	–
Stage-2 (RL)	**53.5** (↑8.4)	**69.1** (↑40.8)	**358**

Cube tasks. As illustrated in Tab. 3, AlloSpatial-4B reaches 69.1% overall accuracy on MindCube, clearly outperforming Gemini-2.5-Pro, Think3D-4B, Spatial-MLLM-4B, and Cambrian-S-3B, with particularly large gains on AROUND and AMONG. In contrast, improvements on precise numerical estimation tasks, such as absolute distance and object counting, are less uniform. This pattern is consistent with the design of World2Mind: the AST and route maps provide robust coarse allocentric structure for relational reasoning, but the current reconstruction pipeline still limits fine-grained metric accuracy. Overall, **AlloSpatial agents are most effective when the task benefits from a stable allocentric organization rather than exact metric reconstruction.**

4.4 Ablation Studies and Additional Results

Effect of input-frame number. We study how AlloSpatial behaves under different observation budgets. Fig. 2 reports performance with 0, 3, 7, 15, and 24 uniformly sampled input frames. AlloSpatial shows the clearest advantage when visual observations are sparse. In the 0-frame setting, where the model relies only on structured AST text, AlloSpatial improves over Qwen3-VL by +18.2 points. With only 3 and 7 frames, it remains consistently stronger than both Qwen3-VL and GPT-5.2, reaching 50.0 and 53.5 overall, respectively. As the number of frames increases, the gap to Qwen3-VL narrows, suggesting that dense visual coverage can partially compensate for missing allocentric memory. This trend indicates that AlloSpatial is particularly effective under limited observations.

Comparison with general thinking mode. We next compare AlloSpatial with the general thinking of Qwen3-VL-4B. As shown in Tab. 4, enabling thinking improves Qwen3-VL-4B on MindCube from 28.3 to 36.1, but brings only a marginal gain on VSI-Bench (45.1 → 45.5) while increasing the average response length to 1064 tokens. In contrast, AlloSpatial-4B after RL reaches 53.5 on VSI-Bench and 69.1 on MindCube with only average 358 tokens. This suggests that the improvement does not come from longer chain-of-thought alone. Instead, the harnessed process provides a more efficient reasoning structure by grounding intermediate reasoning in explicit allocentric evidence.

Comparison with QA-only training. We further perform the same two-stage training schedule on standard QA pairs constructed from the same data, to isolate whether the gains of AlloSpatial come from direct answer supervision or potential benchmark leakage. As shown in Tab. 4, QA-only SFT substantially improves MindCube from 28.3 to 53.9, but decreases VSI-Bench from 45.1 to 43.1. This suggests that answer supervision alone can fit certain benchmark patterns, but does not consistently improve spatial reasoning. QA-only RL slightly improves VSI-Bench to 46.2 and maintains a strong MindCube score of 53.0, yet remains well below AlloSpatial Stage-2, which reaches 53.5 on VSI-Bench and 69.1 on MindCube. This gap indicates that RL over QA supervision alone is insufficient; the main improvement comes from optimizing complete tool-use trajectories that follow the Spatial Reasoning Harness and interact with World2Mind during reasoning.

5 Conclusion & Limitations

We presented **AlloSpatial**, an agentic framework that equips MFMs with allocentric spatial reasoning capability through the proposed World2Mind cognitive mapping sandbox and carefully designed spatial reasoning harness. Experiments on VSI-Bench and MindCube show that AlloSpatial improves proprietary models in a training-free setting and enables compact open-weight agents to outperform larger general-purpose and spatially specialized baselines. A remaining limitation of our study is numerical reasoning: current World2Mind representations provide robust allocentric structure, but reconstruction drift and imperfect metric calibration can still limit precise distance, size, and counting estimates. Future work may address this by incorporating stronger metric calibration and uncertainty-aware reconstruction into the cognitive mapping backend.

Acknowledgement

This work was supported in part by the Project of the National Natural Science Foundation of China under Grant 62576020, 62572467 and in part by the Fundamental Research Funds for the Central Universities.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Jacob LS Bellmund, Peter Gärdenfors, Edvard I Moser, and Christian F Doeller. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415):eaat6766, 2018.

[3] Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 10 (12):551–557, 2006.

[4] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE, 2025.

[5] Meng Cao, Xingyu Li, Xue Liu, Ian Reid, and Xiaodan Liang. Spatialdreamer: Incentivizing spatial reasoning via active mental imagery. *arXiv preprint arXiv:2512.07733*, 2025.

[6] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[8] Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models. *arXiv preprint arXiv:2509.17664*, 2025.

[9] Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. Spacetools: Tool-augmented spatial reasoning via double interactive rl. *arXiv preprint arXiv:2512.04069*, 2025.

[10] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

[11] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408, 2025.

[12] Nianchen Deng, Lixin Gu, Shenglong Ye, Yinan He, Zhe Chen, Songze Li, Haomin Wang, Xingguang Wei, Tianshuo Yang, Min Dou, et al. Internspatial: A comprehensive dataset for spatial reasoning in vision-language models. *arXiv preprint arXiv:2506.18385*, 2025.

[13] Google. Gemini 3.1 pro: Best for complex tasks and bringing creative concepts to life. https://deepmind.google/models/gemini/pro/, 2026.

[14] Google DeepMind. Gemma 4, 2026. URL https://deepmind.google/models/gemma/gemma-4/. Accessed: 2026-05-06.

[15] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024.

[16] Jiaxin Huang, Ziwen Li, Hanlve Zhang, Runnan Chen, Xiao He, Yandong Guo, Wenping Wang, Tongliang Liu, and Mingming Gong. Surprise3d: A dataset for spatial understanding and reasoning in complex 3d scenes. *arXiv preprint arXiv:2507.07781*, 2025.

[17] Shuai Huang, Wenxuan Zhao, and Jun Gao. Si-bench: Benchmarking social intelligence of large language models in human-to-human conversations. *arXiv preprint arXiv:2510.23182*, 2025.

[18] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[19] Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, et al. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025.

[20] Zhanpeng Luo, Ce Zhang, Silong Yong, Cunxi Dai, Qianwei Wang, Haoxi Ran, Guanya Shi, Katia Sycara, and Yaqi Xie. pyspatial: Generating 3d visual programs for zero-shot spatial reasoning. *arXiv preprint arXiv:2603.00905*, 2026.

[21] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*, 2025.

[22] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

[23] Zhenhua Ning, Zhuotao Tian, Shaoshuai Shi, Guangming Lu, Daojing He, Wenjie Pei, and Li Jiang. Enhancing spatial reasoning in multimodal large language models through reasoning-based segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7851–7860, 2025.

[24] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978.

[25] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023.

[26] Kun Ouyang. Spatial-r1: Enhancing mllms in video spatial reasoning. *arXiv e-prints*, pages arXiv–2504, 2025.

[27] Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

[28] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024.

[29] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. *arXiv preprint arXiv:2307.06135*, 2023.

[30] Shouwei Ruan, Liyuan Wang, Caixin Kang, Qihui Zhu, Songming Liu, Xingxing Wei, and Hang Su. From reactive to cognitive: brain-inspired spatial intelligence for embodied agents. *intelligence (AGI)*, 3(9):10, 2025.

[31] Daniela Schiller, Howard Eichenbaum, Elizabeth A Buffalo, Lila Davachi, David J Foster, Stefan Leutgeb, and Charan Ranganath. Memory and space: towards an understanding of the cognitive map. *Journal of Neuroscience*, 35(41):13904–13911, 2015.

[32] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[33] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

[34] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[35] Qineng Wang, Baiqiao Yin, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Mindcube: Spatial mental modeling from limited views. *arXiv e-prints*, pages arXiv–2506, 2025.

[36] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[37] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Yuxin Wang, Lei Ke, Boqiang Zhang, Tianyuan Qu, Hanxun Yu, Zhenpeng Huang, Meng Yu, Dan Xu, and Dong Yu. N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. *arXiv preprint arXiv:2512.16561*, 2025.

[39] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

[40] Mingrui Wu, Zhaozhi Wang, Fangjinhua Wang, Jiaolong Yang, Marc Pollefeys, and Tong Zhang. From indoor to open world: Revealing the spatial reasoning gap in mllms. *arXiv preprint arXiv:2512.19683*, 2025.

[41] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[42] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025.

[43] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[44] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. URL https://arxiv.org/abs/2407.12772.

[45] Weichen Zhang, Ruiying Peng, Chen Gao, Jianjie Fang, Xin Zeng, Kaiyuan Li, Ziyou Wang, Jinqiang Cui, Xin Wang, Xinlei Chen, et al. The point, the vision and the text: Does point cloud boost spatial reasoning of large language models? *arXiv preprint arXiv:2504.04540*, 2025.

[46] Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Zhenfei Yin, et al. Think3d: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*, 2026.

[47] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

[48] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Training & Evaluation Configuration

RL training configuration. AlloSpatial-4B and AlloSpatial-8B are initialized from their corresponding supervised cold-start checkpoints at step 240, after three epochs of SFT. We optimize both agents with GSPO using `ms-swift` framework [47] and DeepSpeed ZeRO-2. The RL prompts are sampled from a shared training pool of 59,981 examples, including 49,981 VSI-style examples from VSI-590K and 10,000 examples from the MindCube training split. Inline validation is performed on 1,442 held-out examples, consisting of 392 VSI-Bench-tiny questions and 1,050 MindCube-tiny questions. For each prompt, the policy samples 8 rollouts, with at most 5 tool-interaction turns per rollout. The maximum sequence length is 32,768 tokens, and generated completions are capped at 8,192 tokens. The trajectory reward follows HGTR, combining structural compliance, answer accuracy, tool-use effectiveness, and length control with weights 0.15, 0.60, 0.10, and 0.15, respectively. Validation rewards are logged for monitoring but are not used for optimization.

Both runs use cosine learning-rate decay with a warmup ratio of 0.005, freeze the vision tower and aligner modules, and adopt $\epsilon_{\text{high}} = 0.28$ with overlong-completion filtering. The training scripts are epoch-based rather than hard-coded with a fixed maximum number of steps; the reported checkpoints are selected by validation performance and inference efficiency. Detailed configuration information is shown in Tab. 5.

Table 5: **GSPO training configuration for the reported AlloSpatial checkpoints.** The main paper reports the 600-step AlloSpatial-4B checkpoint and the 400-step AlloSpatial-8B checkpoint. Unique prompts are counted after grouping the 8 sampled rollouts per prompt.

Setting	**AlloSpatial-4B**	**AlloSpatial-8B**
Initialization	Qwen3-VL-4B SFT step 240	Qwen3-VL-8B SFT step 240
Reported RL checkpoint	step 600	step 400
Trainer processes	4	6
Per-device train batch	4	2
Gradient accumulation	4	4
Generated trajectories/update	64	48
Unique prompts/update	8	6
Rollouts per prompt	8	8
Unique prompts to reported ckpt.	4.8K	2.4K
Learning rate	1×10^{-6}	5×10^{-7}
Precision / attention	bf16 / FlashAttention	bf16 / FlashAttention
Trainable modules	language modules	language modules
Optimizer infrastructure	full tuning + ZeRO-2	full tuning + ZeRO-2
Sampling temperature	1.0	1.0
KL coefficient β	0.01	0.0
Eval / save interval	100 steps	50 steps

Evaluation configuration. For trained local agents, we evaluate AlloSpatial with lmms-eval framework [44]. The evaluator parses final responses from predefined answer tags, uses temperature 1.0, allows up to 8 reasoning turns, and caps each completion at 8,192 new tokens. For VSI-Bench, the main post-trained evaluation follows the limited-observation setting with 7 uniformly sampled frames, while the frame-budget ablation evaluates 0, 3, 7, 15, and 24 input frames. For training-free proprietary-model evaluation, we uniformly sample up to 32 frames to provide sufficient observations for external cognitive mapping. MindCube is evaluated using the provided multi-view images.

Across local-agent experiments, World2Mind uses the same default cognitive mapping pipeline, including monocular depth and pose estimation, SAM3-based open-vocabulary segmentation, confidence filtering, semantic point-cloud construction, AST serialization, and optional top-down map rendering. During evaluation, tool calls generated by the model are executed online, and the returned ASTs, route maps, or visualizations are inserted into the dialogue context for subsequent harness-guided reasoning.

B World2Mind Service Parallelization

RL rollout generation can produce many concurrent World2Mind calls, while a single World2Mind process would serialize reconstruction and mapping requests. We therefore deploy World2Mind as a multi-process HTTP service during training and local evaluation. Each service process is assigned to an independent NPU worker and initializes the same cognitive mapping pipeline, including monocular geometry estimation, SAM3 segmentation, semantic point-cloud construction, AST generation, and route-map rendering. In our main configuration, eight World2Mind workers are launched in parallel to support concurrent tool execution.

Within each worker, NPU-intensive depth estimation and segmentation are executed under a worker-level lock, while downstream mapping, AST construction, and route-map generation proceed after NPU computation and are controlled by CPU-side concurrency limits. This separation prevents concurrent rollouts from over-subscribing NPU memory while allowing lightweight mapping operations to proceed efficiently.

Rollout and evaluation clients dispatch each `cognitive_map` request to an available World2Mind worker using a simple load-balancing strategy based on current in-flight requests. Failed or unavailable workers are skipped and requests are retried on another endpoint when possible. This service design improves throughput and tail-latency stability under concurrent rollout generation, while keeping the World2Mind reconstruction pipeline and AST semantics unchanged.

C Spatial Reasoning Harness Prompts

The Spatial Reasoning Harness is implemented with three coupled components: a system message, a machine-readable tool interface, and a user-side reasoning protocol. For proprietary models, the tool interface is passed through function-calling schemas. For locally trained agents, the same interface is serialized as `<tool_call>` blocks containing a JSON object with `name` and `arguments`. Visual tokens are placed before the reasoning protocol, ensuring that the model observes the frames or multi-view images before receiving step-by-step instructions.

System Message

Role. You are a spatial intelligence assistant that analyzes videos and images to answer spatial questions.

Available tools.

- `world2mind`: builds an allocentric cognitive map and returns per-instance spatial data, including coordinates, sizes, and object relations.
- `view_image`: renders cognitive-map visualizations, including top-down landmark maps, route maps, semantic maps, and point-cloud views.

Reliability note. The cognitive map is produced from monocular reconstruction and may contain missing objects, ghost instances, coordinate drift, or segmentation errors. Treat tool outputs as supplementary evidence rather than ground truth, and cross-validate them against direct visual observations before answering.

Answer format. Wrap the final answer in `<Answer></Answer>` tags.

Tool Interface

`world2mind`**.** Generate a query-conditioned cognitive map from the current visual input. The tool automatically analyzes the video frames or images already provided in the conversation; the model does not provide file paths.

Returned knowledge.

- **Landmark knowledge** (`knowledge_type=landmark` or `both`): detected instances, metric coordinates, footprint size, and object relations.
- **Route knowledge** (`knowledge_type=route` or `both`): traversable regions, camera trajectory, and grid-based route information.

Parameter	Requirement	Meaning
`categories`	required; list of strings	Object categories to detect as landmarks, e.g., `chair`, `table`, `door`.
`scene_type`	required; `indoor` or `outdoor`	Selects scene-specific reconstruction settings.
`knowledge_type`	required; `landmark`, `route`, or `both`	Chooses landmark-only, route-only, or complete cognitive-map output.
`output_format`	required; `rectangle` or `ellipse`	Specifies the landmark footprint representation.
`traversable_categories`	required for `route` or `both`	Ground or surface categories used to infer passable regions, e.g., `floor`, `carpet`, `road`.

`view_image`**.** Inspect a visualization generated by the most recent `world2mind` call. The required parameter `visualization_type` must be selected from the available visualizations returned by `world2mind`, such as `landmark_vis`, `route_vis`, `pointcloud_rgb_topdown`, or `pointcloud_semantic_topdown`. The model should call `view_image` when the YAML output is ambiguous, when object layout requires visual verification, or when map evidence must be checked against raw observations.

Text-Form Tool Calls for Local Agents

For local SFT/GSPO agents, tool calls are serialized as text while preserving the same schema:

```
<tool_call>
{
  "name": "world2mind",
  "arguments": {
    "categories": ["chair", "table", "door"],
```

```
    "scene_type": "indoor",
    "knowledge_type": "both",
    "output_format": "rectangle",
    "traversable_categories": ["floor", "carpet"]
  }
}
</tool_call>
```

The schema requires `categories`, `scene_type`, `knowledge_type`, and `output_format`. Malformed calls are returned as tool-error messages when applicable and are not treated as valid spatial evidence.

User-Side Reasoning Protocol

Visual tokens. The video frames or multi-view images are placed before the textual instructions.

Step 1: Visual Clues. Describe concrete visual observations before any tool call, including visible objects, relative positions, spatial relations, and a preliminary answer when possible.

Step 2.1: `world2mind` Tool Call. Call `world2mind` only when the question requires information that vision alone cannot reliably provide, such as metric distance, 3D coordinates, route layout, viewpoint transformation, or complex spatial relations.

Step 2.2: Map Clues. If `world2mind` is called, summarize map evidence only: detected instances, coordinates, sizes, containment relations, route cells, trajectories, or camera orientations. Do not reconcile it with visual clues yet.

Step 3.1: `view_image` Tool Call. If the structured map is ambiguous or layout verification is needed, call `view_image` using one of the available visualization types returned by `world2mind`.

Step 3.2: Visualization Clues. Describe the rendered cognitive-map visualization, such as top-down layout, object arrangement, route trajectory, or conflicts with the raw visual input.

Step 4: Cross-Validation. Compare visual evidence, map evidence, and visualization evidence. Identify conflicts caused by reconstruction drift, missing objects, false detections, or visual ambiguity, and decide which evidence should dominate.

Step 5: Final Answer. Integrate the validated evidence and provide the conclusion. For single-word, numeric, or option questions, place only the final value inside the answer tag.

Question. `{query}`

D Case Analysis

D.1 Case 1: Metric Closest-Point Reasoning

User Turn: Input Observation and Query

Observation. The input is a video of a kitchen. Eight approximately uniformly sampled frames are shown below as a two-row, four-column contact sheet.



Question. Determine the shortest distance between the towel and the refrigerator in meters. Please answer using a single word or phrase.

Assistant Turn 1: Visual Clues and World2Mind Call

Visual clues. The model first observes a green patterned towel hanging on the oven door and a white refrigerator in other frames. Because the two objects are not visible in a single calibrated view, it estimates that direct visual inspection is insufficient for precise metric distance.

Tool call.

```
<tool_call>
{
  "name": "world2mind",
  "arguments": {
    "categories": ["towel", "refrigerator", "stove", "oven"],
    "knowledge_type": "landmark",
    "output_format": "rectangle",
    "scene_type": "indoor",
    "traversable_categories": ["floor"]
  }
}
</tool_call>
```

Tool Result 1: World2Mind Landmark AST

World2Mind returns a JSON tool result with `success=true`, scene id `scene0220_01_c70da29e-5e7`, available visualization `landmark_vis`, and the following AST in the `landmark_yaml` field.

```
Map_Metadata:
  Coordinate_System: 2D Absolute Grid (meters)
  Origin_Ref: (-3.129, -1.293)
  AST_Format: Rectangle (center, width, height)
Spatial_Hierarchy:
- Node_ID: floor_01
  Spatial_Pose:
    Center: (-0.087, 1.449)
    Width: 4.117
    Height: 2.838
  Attributes:
    Height_Range: (1.179, 1.585)
    Area_m2: 7.1149
    Point_Count: 15383
- Node_ID: refrigerator_01
  Spatial_Pose:
    Center: (0.474, 1.475)
    Width: 0.924
    Height: 0.866
  Attributes:
    Height_Range: (-0.234, 1.525)
    Area_m2: 0.5815
    Point_Count: 2674
- Node_ID: stove_01
  Spatial_Pose:
    Center: (2.106, 0.691)
    Width: 0.65
    Height: 0.731
  Attributes:
    Height_Range: (0.554, 1.585)
    Area_m2: 0.4324
    Point_Count: 3588
  Contains_Children:
  - Node_ID: towel_01
    Spatial_Pose:
      Center: (1.918, 0.634)
      Width: 0.052
      Height: 0.304
    Attributes:
      Height_Range: (0.839, 1.207)
      Area_m2: 0.0102
```

```
      Point_Count: 1136
  Relation_To_Parent: Inside
```

Assistant Turn 2: Map Calculation and Visualization Request

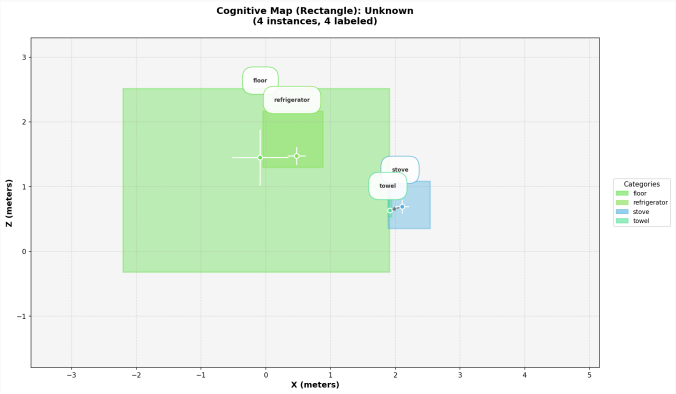
Map clues. The model reads `towel_01` center $(1.918, 0.634)$ m and `refrigerator_01` center $(0.474, 1.475)$ m. The center-to-center distance is about 1.671 m, but the question asks for nearest-point distance. Using rectangle half-extents, the horizontal gap is $1.444 - (0.026 + 0.462) = 0.956$ m and the vertical gap is $0.841 - (0.152 + 0.433) = 0.256$ m. The nearest-boundary distance is therefore $\sqrt{0.956^2 + 0.256^2} \approx 0.990$ m.

Visualization request. The trajectory then requests a map visualization to verify the geometric layout before finalizing the answer.

```
<tool_call>
{
  "name": "view_image",
  "arguments": {"visualization_type": "landmark_vis"}
}
</tool_call>
```

Tool Result 2: Landmark Visualization

The returned `landmark_vis` image provides a top-down check of the refrigerator and stove/towel layout.



Assistant Turn 3: Cross-Validation, Final Answer, and Ground Truth

Visualization clues. The top-down landmark visualization places the refrigerator and stove/towel on different sides of the kitchen, with a gap consistent with the rectangle-derived estimate.

Cross-validation. The raw visual evidence suggests a short, plausible kitchen-scale separation, and the map does not introduce a visible contradiction. The agent therefore trusts the AST-derived nearest-point estimate.

Final answer.

```
Prediction: <Answer>0.99 m</Answer>
Ground truth (VSI): 1.1 m
Absolute error: 0.11 m
```

D.2 Case 2: Viewpoint-Conditioned Navigation

User Turn: Input Observations and Query

Observation. The input consists of four wall-aligned views of the same black chair: front, left, back, and right.



Image 1 Image 2 Image 3 Image 4

Question. If the observer stands at the same spot and faces the same direction as image 4, then turns left and moves forward, will the observer get closer to the grey sofa? A. No B. Yes.

Assistant Turn 1: Visual Clues and World2Mind Call

Visual clues. The model identifies image 1 as the front view of the chair, image 2 as the left view, image 3 as the back view, and image 4 as the right view. The grey sofa is clearly visible in image 3. From image 4, turning left appears to rotate the observer toward the direction where the sofa may lie, but the view transformation is ambiguous enough to benefit from an allocentric map.

Tool call.

```
<tool_call>
{
  "name": "world2mind",
  "arguments": {
    "categories": ["black chair", "grey sofa", "desk", "plant", "bag", "window"],
    "knowledge_type": "both",
    "output_format": "rectangle",
    "scene_type": "indoor",
    "traversable_categories": ["floor", "carpet", "tile flooring"]
  }
}
</tool_call>
```

Tool Result 1: World2Mind AST and Route Knowledge

World2Mind returns both landmark and route knowledge for scene `images_4_816b9d06-732`. The complete AST and route fields are shown below.

AST.

```
Map_Metadata:
  Coordinate_System: 2D Absolute Grid (meters)
  Origin_Ref: (-2.281, -0.401)
  AST_Format: Rectangle (center, width, height)
  Camera_Views:
  - Frame: frame_000000
    Image: image 1
    Position: (-0.975, 0.668)
    Heading_Deg: 56.2
  - Frame: frame_000001
    Image: image 2
    Position: (0.921, 0.676)
    Heading_Deg: -57.6
```

```
  - Frame: frame_000002
    Image: image 3
    Position: (-0.0, 0.0)
    Heading_Deg: 0.3
  - Frame: frame_000003
    Image: image 4
    Position: (-0.17, 1.443)
    Heading_Deg: 161.8
Spatial_Hierarchy:
- Node_ID: grey sofa_01
  Spatial_Pose:
    Center: (1.524, 1.234)
    Width: 0.708
    Height: 1.022
  Attributes:
    Height_Range: (-0.654, 0.881)
    Area_m2: 0.4026
    Point_Count: 640
- Node_ID: black chair_01
  Spatial_Pose:
    Center: (-0.077, 1.314)
    Width: 0.532
    Height: 0.333
  Attributes:
    Height_Range: (-0.529, 0.49)
    Area_m2: 0.1173
    Point_Count: 2911
- Node_ID: bag_01
  Spatial_Pose:
    Center: (0.32, 1.278)
    Width: 0.412
    Height: 0.185
  Attributes:
    Height_Range: (0.311, 0.672)
    Area_m2: 0.0523
    Point_Count: 296
- Node_ID: plant_01
  Spatial_Pose:
    Center: (0.214, 1.863)
    Width: 0.307
    Height: 0.22
  Attributes:
    Height_Range: (-0.187, 0.221)
    Area_m2: 0.0419
    Point_Count: 516
```

Route knowledge.

```
Route_Knowledge_Metadata:
  Coordinate_System: Grid (NxN cells)
  Grid_Divisions: 10
  Cell_Size: (0.474, 0.427)
  Scene_Bounds:
    Min: (-2.281, -0.401)
    Max: (2.458, 3.865)
  Total_Traversable_Cells: 32
  Estimated_Floor_Area_m2: 6.47
  Trajectory_Length: 4
  Note: Estimated_Floor_Area_m2 is the traversable floor area only. Actual room area
    approximately Floor_Area + furniture footprint area. Do NOT use Scene_Bounds to
     calculate
    room area -- Scene_Bounds is the axis-aligned bounding box of the entire 3D
     reconstruction
    and always significantly overestimates room size.
  Description: Route knowledge map showing traversable areas and camera movement path
  Scene: images_4_816b9d06-732
Traversable_Grid:
```

```
- (0, 3)
- (1, 3)
- (1, 4)
- (1, 6)
- (1, 7)
- (1, 8)
- (2, 3)
- (2, 6)
- (2, 7)
- (3, 2)
- (3, 3)
- (3, 4)
- (3, 5)
- (3, 6)
- (3, 7)
- (4, 2)
- (4, 3)
- (4, 4)
- (4, 5)
- (4, 6)
- (4, 7)
- (5, 3)
- (5, 4)
- (5, 5)
- (5, 6)
- (6, 3)
- (6, 4)
- (6, 5)
- (6, 6)
- (7, 3)
- (7, 5)
- (7, 6)
Camera_Trajectory: (2, 2) --> (2, 6) --> (0, 4) --> (4, 4)
Camera_Orientations:
- Frame: frame_000000
  Image: image 1
  Position: (-0.975, 0.668)
  Heading_Deg: 56.2
- Frame: frame_000001
  Image: image 2
  Position: (0.921, 0.676)
  Heading_Deg: -57.6
- Frame: frame_000002
  Image: image 3
  Position: (-0.0, 0.0)
  Heading_Deg: 0.3
- Frame: frame_000003
  Image: image 4
  Position: (-0.17, 1.443)
  Heading_Deg: 161.8
```

Assistant Turn 2: Map Clues and View-Image Call

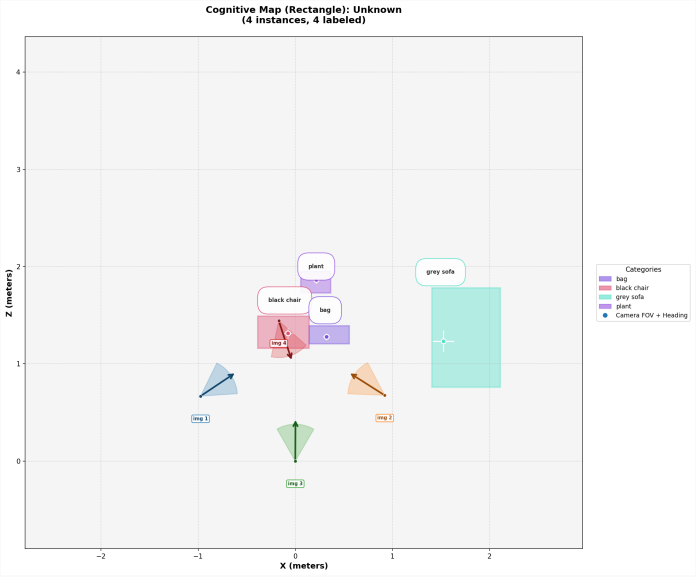
Map clues. The map places the grey sofa at $(1.524, 1.234)$ and the black chair at $(-0.077, 1.314)$. Image 4 has camera position $(-0.170, 1.443)$ and heading $161.8°$. Turning left changes the heading to approximately $71.8°$, which points toward increasing x. Since the sofa has a much larger x coordinate than the image-4 camera, moving forward after the left turn should reduce the distance to the sofa.

Tool call.

```
<tool_call>
{
  "name": "view_image",
  "arguments": {"visualization_type": "landmark_vis"}
}
</tool_call>
```

Tool Result 2: Landmark Visualization

The returned landmark visualization provides an explicit top-down check of the chair, sofa, and image-4 camera pose.



Assistant Turn 3: Visualization Clues, Cross-Validation, Final Answer, and Ground Truth

Visualization clues. The map shows the black chair near the center and the grey sofa to the right/east of the image-4 camera position. This agrees with the coordinate-level reading.

Cross-validation. Visual evidence from image 3 shows the sofa behind the chair, while image 4 is the right-side view. The map evidence shows that a left turn from heading 161.8° points approximately toward heading 71.8°, which moves toward the sofa's larger x coordinate. The visual and allocentric cues therefore agree.

Final answer.

```
Prediction: <Answer>B</Answer>
Ground truth (MindCube): B (Yes)
```

E Computational Costs in Training and Inference

Training cost. In our setup, both AlloSpatial-4B and AlloSpatial-8B are trained on 8 HUAWEI Ascend 910B NPUs, taking approximately 60 and 40 NPU-hours, respectively, to reach the reported checkpoints. The 4B and 8B agents use 4 and 6 trainer processes, respectively. Around the reported checkpoints, one GSPO update typically takes 4–6 minutes, including multi-turn rollout generation, live World2Mind execution, reward computation, and policy optimization. The exact update time varies with the tool-use rate, the number of sampled World2Mind calls, and the reconstruction difficulty of each rollout batch.

Inference cost. Inference cost is determined primarily by the number of input frames and the number of World2Mind calls. Each AlloSpatial query may involve multiple reasoning turns, and each valid World2Mind call can trigger depth and pose estimation, segmentation, semantic alignment, AST construction, and optional route-map rendering. With the parallel World2Mind service, VSI-Bench-tiny evaluation takes roughly 12 minutes for 392 questions with concurrent evaluation and 8 World2Mind workers.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions: AlloSpatial, World2Mind, the Spatial Reasoning Harness, and the RL-based internalization into open-weight agents. The claims are supported by evaluations on VSI-Bench and MindCube, including training-free proprietary-model results, trained Qwen3-VL-based agents, ablations, and limitations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper discusses limitations in the Conclusion and Limitations section, especially the remaining weakness on fine-grained numerical spatial reasoning due to reconstruction drift and imperfect metric calibration. The appendix also reports computational costs and implementation constraints of online World2Mind execution.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical theorems or formal proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the benchmarks, data splits, metrics, frame-sampling protocols, baselines, training stages, reward design, evaluation settings, prompt format, and World2Mind service configuration. Additional hyperparameters, computational costs, and case analyses are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper uses public benchmarks and datasets where available and provides detailed implementation, training, and evaluation settings in the appendix. Code and checkpoints are not included in the anonymous submission; we plan to release reproducibility materials subject to licensing and anonymization constraints.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main paper specifies datasets, benchmarks, metrics, frame budgets, baselines, model variants, and the two-stage training procedure. The appendix provides GSPO training configurations, reward weights, rollout settings, evaluation parameters, World2Mind parallelization, and computational costs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports benchmark scores on official Tiny subsets and controlled ablations, but does not report error bars or confidence intervals. Repeating full RL training and proprietary-model evaluations multiple times would be computationally expensive; instead, the paper provides task-level breakdowns, ablations, and comparisons across multiple model families.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the training infrastructure, including 8 NVIDIA H200 GPUs, trainer process counts, GSPO update time, approximate wall-clock training time for AlloSpatial-4B and AlloSpatial-8B, World2Mind service workers, and inference throughput.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses public benchmarks, existing foundation models, and generated tool-use trajectories for spatial reasoning research. It does not involve deception, human-subject experimentation, private personal data collection, or unsafe data release.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses AlloSpatial as foundational research toward more spatially capable multimodal and embodied agents. It also notes that incorrect spatial reasoning or over-trusting reconstructed maps can be harmful in safety-critical settings, motivating the harness-based cross-validation design and the limitation discussion on metric reliability.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce a scraped dataset, image generator, or high-risk pretrained foundation model release. The proposed method is evaluated as a spatial reasoning framework, and any future release of code or checkpoints will follow the licenses and usage restrictions of the underlying models and datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the existing datasets, benchmarks, foundation models, geometry models, segmentation models, and spatial reasoning baselines used in the study. We use these assets for research evaluation and training under their respective terms and do not redistribute restricted proprietary models or datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces AlloSpatial, World2Mind, the Spatial Reasoning Harness, and trained AlloSpatial agents. The method, prompts, tool schema, reward design, training configuration, evaluation setup, and service parallelization are documented in the main paper and appendix; release of code or checkpoints will follow anonymization and licensing constraints.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All evaluations are performed on existing benchmarks and generated model trajectories.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject research, user studies, or collection of participant data. Therefore, IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs and MFMs are core components of the research. The paper describes the use of proprietary models for training-free evaluation and trajectory distillation, Qwen3-VL as the open-weight backbone for AlloSpatial agents, and LLM-based multi-turn tool-use reasoning as part of the proposed method.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.